Applied Data Analysis CS-401, Project

Reda El Alaoui, Hugo Rousseau, Anshul Toshniwal
School of Computer and Communication Sciences, EPFL, Switzerland

***Abstract*—In the paper on comparing Random forest with logistic regression for prediction civil war onset [1], the authors state the superiority of Random Forest over logistic regression and its variants. To further extend their analysis, we will consider various other methods like SVM, Least square regression and neural networks and compare them with Random Forest.**

I. Introduction

There are several Machine learning algorithms for classification and it is important to know whether methods other than Random Forest or Logistic regression have a better predictive power on the Civil war Data set. Predicting the onset of civil war is an important Social Science problem and can save many lives if interventions are made at the right time for certain policies.

Onset of civil war is a rare event resulting in the data set being used to describe it being inherently unbalanced. As the data set is imbalanced, it is not appropriate to judge the accuracy of any classifier based on it's prediction accuracy of the civil war, [1] but rather each model is judged on the basis of the ROC (Receiver Operating Characteristic) curve that is generated as the threshold of classification is varied. ROC curves are preferable as they measure the False positive rate or True positive rate rather than the total number of True or False positives predicted by the model which is very biased in an unbalanced data set. ROC curves can be summarised with a single number, the area under the curve(AUC). AUC captures the accuracy of the classifier and ranges between [0,1]. Generally, for a model, AUC value of 0.7 is considered as an average prediction power while a value above 0.9 is correlated with a superior prediction power.

II. Data set

A. Data set description

As we are going to compare various methods, we have chosen to analyse the same data provided in the paper for comparing Random Forest and Logistic Regression. This data set is taken from the Civil War Data (CWD) and contains nearly 286 prediction variables. Each row corresponds to a country-year and contains the values of the corresponding 286 variables believed to influence the onset of civil war. Occurence of a civil war is recorded with a binary variable labelled as 'warstds' with it's value being 1 if a civil war onset was recorded and 0 otherwise. There seem to be no missing values in the data set and thus no pre-processing to handle them is needed.

In this study, we will keep only 90 of these variables. These variables are the same variables used by the authors in the models described by them in the paper and the similar choice of variables will thus help in comparing the models.

III. Ordinary Least Square method

A. Objectives

As the data set is composed of continuous and categorical variables, it seems that the least square method is not the most suitable method. Thus, it seemed interesting to analyze the behavior of this method in front of this non-linear dataset. Indeed, it is legitimate to wonder to what extent the least square method is wrong : Can it, in spite of everything, give coherent results ?
In order to carry out this first study of the least square model. We are going to start by naively carrying out a regression with the **90 initial variables**. Then, we will be able to evaluate the most influential data concerning this model. Indeed, the **p-value** of the variables will allow us to establish the list of variables that allow the best prediction. Finally, we will perform a new training of the model with the significant variables.

B. Basic dataset

In this first step, the 90 variables are used and then the model is trained with the `LinearRegression` function. And, in this first example, we obtain a AUC score of 80%, a surprisingly high score which may be indicative of an approximately linear relationship between feature vectors and the output.

In a binary classification problem, we are interested in the probability of occurrence of an event given the values of a feature vector. Probability, by definition, ranges between 0 and 1. But, with `y_pred`, we can observe that in linear regression, we are predicting an absolute number, which can range outside 0 and 1. Linear regression, in addition to the predictions, provides us error estimates which help us in ranking features

C. Features selection

In this second step, the `statmodel` library is used in order to have a good visualization of the importance of the data. Indeed, `statmodel` allows us to obtain the p-values of each variable, as well as the coefficients of these variables. We can then select the variables that best train the model. Thus, it is observed that of the 90 initial variables, only **16**

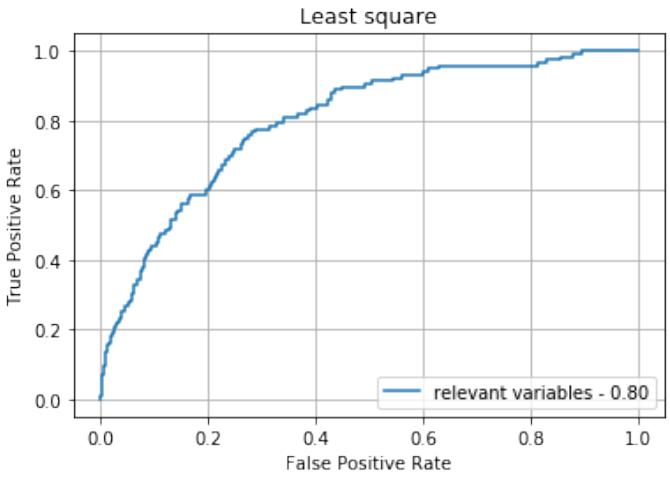

Figure 1. ROC curve and AUC score for Linear Regression

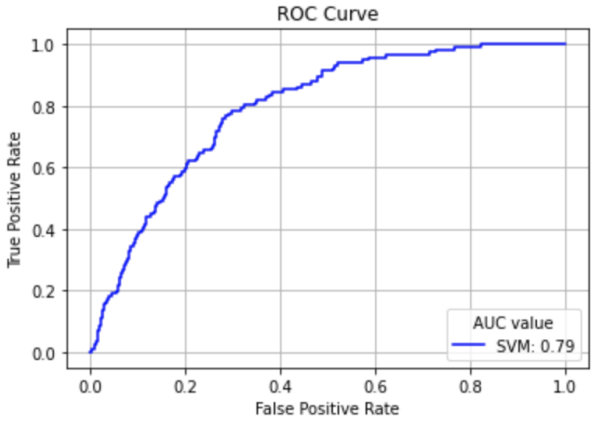

Figure 2. ROC curve and AUC score for SVM

are significant. In fact, based on the P-values obtained with the model created with the `statmodel` library, only 16 variables with a **p-value lower than 5%** are included.

D. New dataset to improve the model ?

In this last step, the model is re-trained with the 16 variables selected in the previous section. The ROC curve obtained is quite similar to the one obtained in step 1. Indeed, they both have an AUC score of 0.80.

E. Intermediate conclusion

Applying linear regression for classification is not an absurd idea, but we observe that the results not quite comparable to the results obtained with logistic regression or other classification methods, which are preferred over linear regression. It would therefore be more interesting to analyze other classification methods that presumably should be more efficient than the least square method.

IV. Support Vector Machines method

In this part, we try to fit a SVM to the binary data to get an understanding of whether this method predicts the onset of civil war more accurately. SVM are known to work quite well in the case of a linearly separated data motivating us to consider this method for analysis. To fit a SVM to the binary data, we use a RBF kernel and fit the 'warstds' column to all the feature vectors. The ROC curves are evaluated by employing 5 fold cross validation. We observes the AUC to be 0.79 which is close to the value obtained by Logistic regression and least square but still less than what was obtained from those methods. However, it is impossible to extract feature imprtance using SVM with RBF kernel which is a disadvantage of using this method.

V. Neural Networks method

Neural networks have been known to have a high predictive accuracy on very diverse datasets and are standard for capturing the non linear dependence between inputs and outputs. We employ a Multi layer perceptron on the dataset to classify the binary data. As the model contains several hyperparameters like number of layers, regularisation strength, activation function and the learning rates, we do a grid search through the library 'GridSearchCV' to find the optimum parameters. After choosing the optimum parameters, we train a multilayer perceptron on the whole dataset to get the ROC curves.

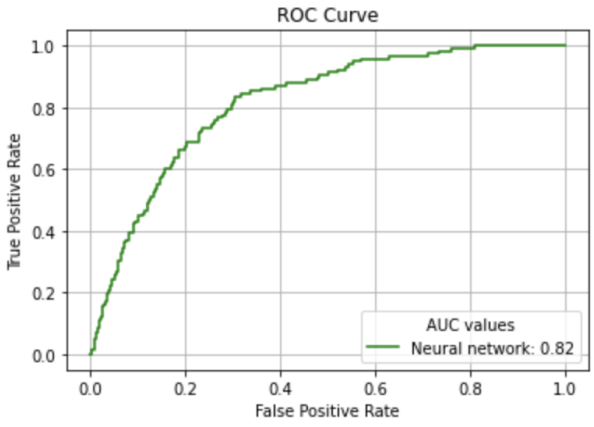

Figure 3. ROC curve and AUC score for Neural networks

The AUC value we obtain from the neural network by choosing the specific hyperparameters is 0.82 which makes this classifier "good". This value may be increased by choosing a broader set for hyperparameter optimisation and by having more data or a more developed dataset.

VI. Bayes Classifier method

In addition to all the above methods, we apply the naive Bayesian classification to benchmark the values obtained by these methods. Bayesian estimator assumes that the value of one feature is independent of another feature given the class variable value. A Gaussian Bayes model is fit to the data and the ROC curve is obtained by 5 fold cross validation. The

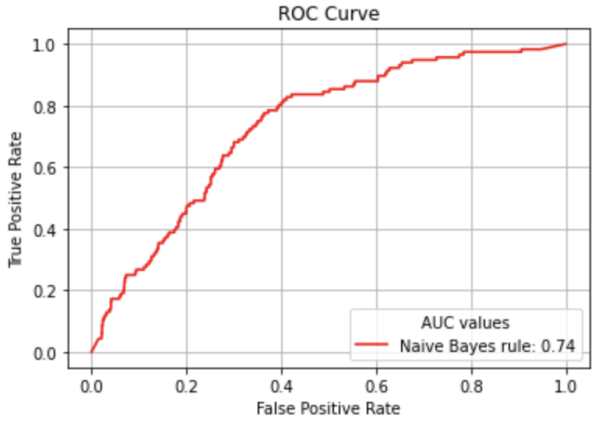

Figure 4. ROC curve and AUC score for Naive bayes method

AUC value that we get is 0.74 which is perhaps the least among all methods and is in line with our expectations. If any method has an AUC lower than 0.74, we should not employ that model in practice as even a straightforward procedure of probability classification can perform better.

VII. Comparison

As it is evident from Figure 5, Random forest remains the most accurate method in terms of AUC. The less accurate method is the Naive Bayesian classification model which simply classifies the data on probability estimates but we should notice that even such a simplistic method provides us at least an average classifier, considering its AUC value. Logistic Regression and its variants, seem to be almost as accurate as Linear Regression. Higher than expected accuracy of Linear Regression implies an approximate linear relationship between variables and the output. The most important features selected by Linear Regression are quite similar to the ones that were identified by a Random Forest model by the authors. Linear Regression identified 16 features among 90 as relevant to the prediction of onset of civil war. Many features like GDP per capita, Export percentage, population density etc were also selected by Random Forest. By observing the coefficients which Linear regression learned, we can interpret a positive or negative correlation of the feature vector on the onset of civil war. For instance, GDP per capita feature had the learned coefficient as -0.114 which is consistent with the observation that countries with higher GDP per capita rarely witness a civil war.

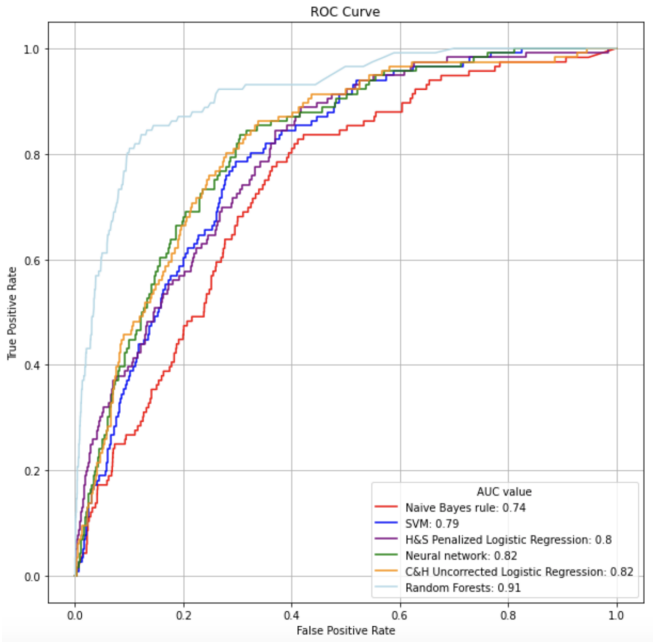

Figure 5. ROC curve and AUC scors for all the methods

VIII. Conclusion

We observe that the best performance is obtained from Random Forest among all the models tried further supporting the ideas presented in the paper about the usefulness of Random Forest in predicting the onset of civil war. Random forests, in addition, provide us the importance of each feature further helping the Social scientists in making causations. The features identified by Linear regression and Random Forest seem similar thus providing confirmation of their importance from 2 different methods. If the relationship between response and predictor variables was linear, Linear Regression would have provided a better estimate, but, we observe that Random Forest has a higher AUC implying that the relationship between the response and predictor variables can not be completely captured with a linear relationship. Hence, it makes sense to initially model a class imbalanced data with Random Forest to gain more insights about the feature importance and have good prediction accuracy. Finally, the good AUC score of Neural Network made us think that it can be another lead in order to develop a civil war onset predictor. For the moment, Random Forest is still the best way to predict civil war onsets and is a powerful starting point in political sciences. However, Neural Networks in political data science can also be seen as an important field to develop and to improve.

References

[1] D. Muchlinski, D. Siroky, J. He, and M. Kocher, "Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data," *Political Analysis*, vol. 24, pp. 1–17, 11 2015.